DOCUMENT RESUME

ED 226 018 TM 830 034

AUTHOR Wilcox, Rand R.
TITLE R. & D. in Psychometrics: Technical Reports on Latent Structure Models.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Nov 82
GRANT NIE-G-80-0112
NOTE 60p.; Methodology Project.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Educational Testing; Evaluation Methods; *Guessing (Tests); *Latent Trait Theory; Mathematical Models; Multiple Choice Tests; Probability; *Psychometrics; *Research Methodology; Responses; Statistical Analysis; *Test Reliability
IDENTIFIERS *Answer Until Correct; Item Hierarchies

ABSTRACT

This document contains three papers from the Methodology Project of the Center for the Study of Evaluation. Methods for characterizing test accuracy are reported in the first two papers. "Bounds on the K Out of N Reliability of a Test, and an Exact Test for Hierarchically Related Items" describes and illustrates how an extension of a latent structure model can be used in conjunction with results in Sathe (1980) to estimate the upper and lower bounds of the probability of making at least k correct decisions. "An Approximation of the K Out [of] N Reliability of a Test, and a Scoring Procedure for Determining Which items an Examinee Knows" proposes a probability approximation that can be estimated with an answer-until-correct (AUC) test. "How Do Examinees Behave When Taking Multiple Choice Tests?" deals with empirical studies of AUC assumptions. Over two hundred examinees were asked to record the order in which they chose their responses. Findings indicate that Horst's (1933) assumption that examinees eliminate as many distractors as possible and guess at random from among those that remain appears to be a tolerable approximation of reality in most cases. (Author/PN)

Deliverable - November 1982

METHODOLOGY PROJECT

R&D IN PSYCHOMETRICS: TECHNICAL REPORTS ON LATENT STRUCTURE MODELS

Rand R. Wilcox
Study Director

Grant Number
NIE-G-80-0112, P3

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

TECHNICAL REPORTS ON METHODS FOR CHARACTERIZING TEST ACCURACY

Bounds on the K Out of N Reliability of a Test, and an Exact Test for Hierarchically Related Items

An Approximation of the K Out N Reliability of a Test, and a Scoring Procedure for Determining which Items an Examinee Knows

TECHNICAL REPORT ON EMPIRICAL STUDIES OF ANSWER-UNTIL-CORRECT ASSUMPTIONS

How do Examinees Behave When Taking Multiple Choice Tests?

BOUNDS ON THE K OUT OF N RELIABILITY OF A TEST, AND AN EXACT TEST FOR HIERARCHICALLY RELATED ITEMS

Rand R. Wilcox

Center for the Study of Evaluation
University of California, Los Angeles

ABSTRACT

Consider an n-item multiple choice test where it is decided that an examinee knows the answer if and only if he/she gives the correct response. The k out of n reliability of the test, ρ_k, is defined to be the probability that for a randomly sampled examinee, at least k correct decisions are made about whether the examinee knows the answer to an item. The paper describes and illustrates how an extension of a recently proposed latent structure model can be used in conjunction with results in Sathe et al. (1980) to estimate upper and lower bounds on ρ_k. A method of empirically checking the model is discussed.

Consider a randomly sampled examinee responding to a multiple-choice test item. In mental test theory there are, of course, many procedures that might be used to analyze this item. One approach might be as follows. Suppose a conventional scoring procedure is used where it is decided that an examinee knows the correct response if the correct alternative is chosen, and that otherwise the examinee does not know. If it were possible to estimate the probability, τ, of correctly determining an examinee's latent state (whether he/she knows the correct response) based on the above decision rule, this would give an indication of how well the distractors are performing for the typical examinee. The obvious problem is that under normal circumstances, there is no way of estimating this probability unless additional assumptions are made. One approach is to assume that examinees guess at random among the alternatives when they do not know the answer. If this knowledge or random guessing model holds, τ is easily estimated. However, empirical investigations (Bliss, 1980; Cross & Frary, 1977) suggest that this assumption will frequently be violated, and some related empirical results (Wilcox, 1982, in press a) indicate that such a model can be entirely unsatisfactory for other reasons as well.

Another approach is to use a latent structure model, and many such models have been proposed for measuring achievement (e.g., Brownless & Keats, 1958; Marks & Noll, 1967; Knapp, 1977; Dayton & Macready, 1976, 1980; Macready & Dayton, 1977; Wilcox, 1977a, 1977b, 1981a; Bergan et al., 1980). The choice of a model depends on what one is willing to assume in a particular situation. These models make it possible to estimate errors at the item level such as

β = Pr(randomly selected examinee gives the correct response|examinee does not know) [1]

which in turn yields an estimate of τ. An illustration is given in a later section. (For a review of latent structure models vis-a-vis criterion-referenced tests, see Macready and Dayton, 1980.) For some recent general comments on using latent structure models to measure achievement, see Molenaar (1981) and Wilcox (1981b).

Assume for a moment that for each item on an n-item test, an estimate of τ can be made. Let $x_i = 1$ if a randomly selected examinee's latent state is correctly determined for the ith item; otherwise $x_i = 0$. Then $E(x_i) = \tau_i$ ($i = 1, \ldots, n$) is the probability of a correct decision on the ith item where the expectation is taken over the population of examinees.

Within the framework just described, how should an n-item test be characterized? Observing that Σx_i is the number of correct decisions among the n items, an obvious approach is to use

$$\mu = E(\Sigma x_i) = \Sigma \tau_i \qquad [2]$$

where the expectation is over some particular population of examinees. The parameter μ is just the expected number of correct decisions among the n items for a typical examinee.

Knowing μ might not be important for certain types of tests, but surely it is important for some achievement tests. However, even if μ is known exactly, it would be helpful to have some additional related information about Σx_i. For instance, a test constructor would have a better idea of how the test performs if $VAR(\Sigma x_i)$ could be determined. The problem is that $VAR(\Sigma x_i)$ depends on $COV(x_i, x_j)$, but this last quantity is not known, and at present there is no way of estimating it. An alternative approach is to use the k out of n reliability of the test (Wilcox, in press a) which is given by

$$\rho_k = Pr(\Sigma x_i \geq k) \quad . \qquad [3]$$

In other words, if the goal of a test is to determine which of n items an examinee knows, and if a conventional scoring procedure is used, ρ_k is the probability of making at least k correct decisions for the typical examinee.

Suppose, for example, n = 10 and μ is estimated to be 7. Thus, the expected number of correct decisions is 7, but there is no information about the likelihood that at least 7 correct decisions will be made. If ρ_k were known, a test constructor would have some additional and useful information for judging the accuracy of the test. ρ_k might also be used as follows. Suppose it is desired to have $\rho_8 \geq .9$. If μ is estimated to be 9.1, this is encouraging, but it is not clear what implications this has in terms of making at least 8 correct decisions for the typical examinee.

If x_i is independent of x_j, $i \neq j$, an exact expression for ρ_k is available via the compound binomial distribution. Perhaps there are situations where this independence might be assumed, but it is evident that this independence will not always hold. If it can be assumed that $COV(x_i, x_j) \geq 0$, bounds on ρ_k are available (Wilcox, in press). Recently Sathe, Pradhan, and Shah (1980) derived bounds on ρ_k that make no

assumption about $COV(x_i, x_j)$. The main point of this paper is that these bounds can be estimated using an extension of an answer-until- correct (AUC) scoring procedure proposed by Wilcox (1981a).

An Extension of an Answer-Until-Correct Scoring Procedure

As just indicated, an extension of results in Wilcox (1981a) is needed in order to apply the bounds derived by Sathe et al. (1980). First, however, it is helpful to briefly review the procedure and basic assumptions in Wilcox (1981a).

Consider a specific test item having t alternatives from which to choose, one of which is the correct response. Assume examinees respond according to an AUC scoring procedure. This means that examinees choose an alternative, and they are told immediately whether the correct response has been identified. If they are incorrect another response is chosen, and this process continues until they are successful. Special forms are generally available for administering AUC tests which make these tests easy to use in the classroom.

Let ς_{t-1} be the proportion of examinees who know the correct response, and let ς_i (i = 0, ..., t-2) be the proportion of examinees who can eliminate i distractors given that they do not know. Wilcox (1981a) assumes that examinees eliminate as many distractors as they can, and then choose at random from among those that remain. If p_i

is the probability of choosing the correct response on the ith attempt, then

$$p_i = \sum_{j=0}^{t-i} \varsigma_j/(t - j) \quad (i=1, \ldots, t). \qquad [4]$$

Note that the model assumes that at least one effective distractor is being used. Put another way, no distinction is made between examinees who know the answer and examinees who can eliminate all of the distractors. Also, the model assumes Pr (incorrect response|examinee knows) = 0. In certain special cases this assumption can be avoided (e.g., Macready & Dayton, 1977), and the results reported here are easily extended to this case (cf. Molenaar, 1981; Wilcox, 1981b).

Assuming the model holds,

$$\varsigma_{t-1} = p_1 - p_2 \qquad [5]$$

and

$$\tau = \varsigma_{t-1} + 1 - p_1 = 1 - p_2. \qquad [6]$$

If in a random sample of N examinees, y_i examiness are correct on their ith attempt, $\hat{p}_i = y_i/N$ is an unbiased estimate of p_i which yields an estimate of ς_{t-1} and τ.

Although empirical studies suggest that this model will frequently be reasonable (Wilcox, 1982a, 1982b), there are instances where this will not be the case. For example, some items might require a misinformation model, and an appropriate modification of the AUC scoring procedure has been proposed (Wilcox, 1982). The results outlined here are readily extended to this case, and a brief outline of how this can be done is given below.

Consider any two items on an n-item test, say items i and j. Applying results in Sathe et al. requires an estimate of $\tau_{ij}=Pr(x_i=1,x_j=1)$, i.e., the joint probability of making a correct decision for both items i and j. The remainder of this section outlines how this might be done.

It is assumed that an examinee's guessing rate is independent over the items that he/she does not know. This means, for example, that if an examinee can eliminate all but 2 alternatives on item i, and all but 3 alternatives on item j, the probability of choosing the correct response on the first attempt of both items is (1/2)(1/3) = 1/6.

For the two items under consideration, let p_{km} (k, m = 1, ..., t) be the probability that a randomly selected examinee chooses the correct response on the kth attempt of the first item, and the correct response on the mth attempt of the second. If ς_{gh} is the proportion of examinees who can eliminate g distractors from the first item and h distractors from the second (g, h = 1, ..., t-1), then

$$p_{km} = \sum_{i=0}^{t-k} \sum_{j=0}^{t-m} \varsigma_{ij}/[(t - i)(t - j)] \qquad [7]$$

The last expression can be used to express $\zeta_{t-1,t-1}$ in terms of the p_{km}'s which can be used to estimate $\zeta_{t-1,t-1}$. Note that if the first item has t' alternatives, t' ≠ t, simply replace t-k with t'-k in equation 7.

To clarify matters, consider the special case t = 3. Equation 7 says that

$$p_{11} = \zeta_{22} + \zeta_{21}/2 + \zeta_{20}/3 + \zeta_{12}/2 + \zeta_{11}/4 + \zeta_{10}/6 + \zeta_{02}/3 + \zeta_{01}/6 + \zeta_{00}/9 \quad [8]$$

$$p_{12} = \zeta_{21}/2 + \zeta_{20}/3 + \zeta_{11}/4 + \zeta_{10}/6 + \zeta_{01}/6 + \zeta_{00}/9 \quad [9]$$

$$p_{13} = \zeta_{20}/3 + \zeta_{10}/6 + \zeta_{00}/9 \quad [10]$$

$$p_{21} = \zeta_{12}/2 + \zeta_{02}/3 + \zeta_{11}/4 + \zeta_{01}/6 + \zeta_{10}/6 + \zeta_{00}/9 \quad [11]$$

$$p_{22} = \zeta_{11}/4 + \zeta_{10}/6 + \zeta_{01}/6 + \zeta_{00}/9 \quad [12]$$

$$p_{23} = \zeta_{10}/6 + \zeta_{00}/9 \quad [13]$$

$$p_{31} = \zeta_{02}/3 + \zeta_{01}/6 + \zeta_{00}/9 \quad [14]$$

$$p_{32} = \zeta_{01}/6 + \zeta_{00}/9 \quad [15]$$

$$p_{33} = \zeta_{00}/9 \quad [16]$$

Thus, starting with equation 16

$$\zeta_{00} = 9p_{33} \quad [17]$$

$$\zeta_{01} = 6(p_{32} - p_{33}) \quad [18]$$

and eventually ζ_{22} can be expressed in terms of the p_{km}'s. Replacing the p_{km}'s with their usual unbiased estimate yields an estimate of ζ_{22} say $\hat{\zeta}_{22}$. But it can be seen that for the two items under consideration

(items i and j),

$$\tau_{ij} = \zeta_{22} + 1 - p_{11} \qquad [19]$$

Replacing ζ_{22} and p_{11} with $\hat{\zeta}_{22}$ and $\hat{p}_{11}$ yields an estimate of $\tau_{ij} = Pr(x_i=1, x_j=1)$, say $\hat{\tau}_{ij}$. For arbitrary t, τ_{ij} is given by equation 19 with ζ_{22} replaced with $\zeta_{t-1,t-1}$. Note however, that the model implies that certain inequalities among the p_{km}'s must hold. For example, $p_{31} \geq p_{32} \geq p_{33}$. Estimating the p_{km}'s assuming these inequalities are true requires an application of the pool-adjacent violators algorithm (Barlow et al., 1972). Testing these inequalities can be accomplished by applying results in Robertson (1978).

Bounds on p_k

This section describes how the results in the previous section can be used to estimate bounds on p_k. First, however, results in Sathe et al. (1980) are summarized.

Recall that $\mu = \Sigma\tau_i$ and let

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \tau_{ij} \qquad [20]$$

$$U_k = \mu - k \qquad [21]$$

and

$$V_k = (2S - k(k-1))/2 \qquad [22]$$

Then,

$$\rho_k \geq \frac{2V_{k-1} - (k-2)U_{k-1}}{n(n-k+1)} \qquad [23]$$

If $2V_{k-1} < (n+k-2)U_{k-1}$, then

$$\rho_k \geq \frac{2((k^*-1)U_{k-1} - V_{k+1})}{(k^*-k)(k^*-k+1)} \qquad [24]$$

where $k^* + k - 3$ is the largest integer in $2V_{k-1}/U_{k-1}$. Two upper bounds on ρ_k are also given. The first is

$$\rho_k \leq 1 + ((n + k - 1)U_k - 2V_k)/kn \qquad [25]$$

and the second is that if $2V_k < (k - 1)U_k$,

$$\rho_k \leq 1 - 2\,\frac{(k^* - 1)U_k - V_k}{(k - k^*)(k - k^* + 1)} \qquad [26]$$

where $k^* + k - 1$ is the largest integer in $2V_k/U_k$.

An Illustration

To illustrate how ρ_k might be applied and interpreted, observations on seven items were analyzed according to the procedure outlined above. Each item had two distractors, and they were found to be consistent with the assumptions of the answer-until-correct scoring model. (See Wilcox, 1981a). Table 1 shows the observed frequencies for the first two items. The question to be answered is if these seven items are taken to be the whole test, do they give reasonably accurate information about what the typical examinee knows?

As previously mentioned, the model described above implies that various inequalities among the p_{ij}'s must hold. These inequalities were tested at the .25 level of significance with the procedure in Robertson (1978). In every case the observed responses were consistent with the model.

Generally, when estimating ζ_{22} there is no need to estimate all of the ζ's in equations 8-16. For the situation at hand, ζ_{22} can be estimated as follows. First compute

$$\hat{\zeta}_{02'3} = \hat{p}_{31} - \hat{p}_{32} \qquad [27]$$

for the data in Table 1, this is .107. Next compute

$$\hat{\zeta}_{12/2} = \hat{p}_{21} - \hat{p}_{22} - \hat{\zeta}_{02/3} \qquad [28]$$

which is .074. Then

$$\hat{\zeta}_{22} = \hat{p}_{11} - \hat{p}_{12} - \hat{\zeta}_{12/2} - \hat{\zeta}_{02/3} \qquad [29]$$

which is equal to .225. Substituting these values into equation 19, the estimate of τ_{12} is $\hat{\tau}_{12} = .75$. Applying equation 6 to all seven items, it is seen that $\mu = 5.434$. In other words, it is estimated that the expected number of correct decisions is 5.434.

Next consider ρ_5. The value of S was estimated to be 16.929. From equations 20 - 26, this implies that

$$.42 \leq \rho_5 \leq .74. \qquad [30]$$

This analysis suggests that these seven items, taken as a whole, are not very accurate since there is at least a 26 percent chance of making an incorrect decision on three or more items. How should the test be modified? Another important question is to what extent can it be improved? One approach to improving the test is to increase the number of distractors, and another approach is to try to modify or replace the distractors that are being used. The latter approach will be considered first.

The initial step in trying to decide whether to replace or modify the existing distractors is to determine the extent to which they can be improved. This can be done with the Δ measure in Wilcox (1981, eq..20). This measure is just the difference between the maximum possible value of τ and the estimated value given that $\zeta_2 = \hat{\zeta}_2$. Another related measure is the entropy function (see Wilcox, 1981a). This measures the effectiveness of the distractors among the examinees who do not know the correct response by indicating the extent to which $p_2, \ldots, p_t$ are unequal. The closer they are to being equal, the more effective are the distractors, i.e., guessing is closer to being random. It has been pointed out (Wilcox, 1981a) that Δ might be objectionable as a measure of the extent to which $p_2, \ldots, p_t$ are equal, but for present purposes it would seem to be of interest because increasing p_k depends on the extent to which τ can be increased for each item.

Referring to Wilcox (1981a), a little algebra shows that for the case $t = 3$,

$$\Delta = (p_2 - p_3)/2 . \qquad [30]^{3/}$$

For item 1 in Table 1, $\Delta = .024$, and for item 2 it is .034 (Δ is assumed to be positive; so if $p_2 < p_3$, apply the pool-adjacent violator algorithm in which case Δ is estimated to be zero.)

If the number of alternatives for item 1 is increased to $t = 5$, and if guessing is at random, then the value of τ would be .893 which represents an increase of .126 over the value of τ using the existing distractors. Thus, it would seem that one approach to improving item 1 is to find two more distractors that are about as effective as the two being used. Of course in practice, this might be very difficult to do.

Estimating τ_{ij} When There Is Misinformation

Among the 30 items analyzed by Wilcox (in press, a), the observed test scores suggest that two of the items do not conform well to the AUC scoring model described in a previous section. Thus, the proposed estimate of τ_{ij} is inappropriate. This section outlines how this problem might be solved when a misinformation model appears to be more appropriate for some of the items on the test.

Consider a test item with t alternatives, and let ς_t be the proportion of examinees who eliminate the correct response from consideration on their first attempt of the item. (An AUC scoring procedure is being assumed.) Once an examinee eliminates all of the distractors that are consistent with his/her misinformation, it is assumed that the examinee chooses the correct response on the next attempt. This assumption is made here because it seems to give a good approximation to how examinees were behaving on the items used in Wilcox (in press a). It is also assumed that if an examinee does not know and does not have misinformation, then he/she guesses at random among the t alternatives. Finally, for examinees with misinformation, assume that they believe the correct response is one of c alternatives that are in actuality incorrect. Thus, examinees with misinformation will require at least c + 1 attempts before getting the item correct. As an illustration, consider t = 5 and c = 3. Then,

$$p_1 = \zeta_{t-1} + \zeta_{t+1}/5 \quad [32]$$

$$p_2 = \zeta_{t+1}/5 \quad [33]$$

$$p_3 = \zeta_{t+1}/5 \quad [34]$$

$$p_4 = \zeta_t + \zeta_{t+1}/5 \quad [35]$$

$$p_5 = \zeta_{t+1}/5 \quad [36]$$

where ζ_{t+1} is the proportion of examinees who do not know and who do not have misinformation.

Various modifications of the model are, of course, possible and presumably this model (with some appropriately chosen c value) will give a good fit to the observed test scores. For illustrative purposes, equations 32 - 36 are assumed. The point of this section is that it is now possible to again estimate τ_{ij} where the misinformation model is assumed to hold for one or both of the items in any item pair. Note that for a single item where equations 32 - 36 hold,

$$\tau = \zeta_{t-1} + \zeta_{t+1}/t \quad . \quad [45]$$

37

To estimate τ_{ij}, the joint probability of making a correct decision on a pair of items where, say, the first item is represented by a misinformation model, equation 7 must be rederived. Accordingly, let t' be the number of alternatives on the first item, and t is the number of alternatives on the second. The misinformation model assumes that on the first attempt of the item, examinees belong to one of three mutually

exclusive categories, namely, they know the answer and choose it, they have misinformation and eliminate the correct response, or they do not know and guess at random. Thus, using previously established notation, equation 8 becomes,

$$p_{11} = \varsigma_{42} + \varsigma_{41}/2t' + \varsigma_{40}/3t' + \varsigma_{02}/t' + \varsigma_{01}/2t' + \varsigma_{00}/3t' \qquad [38]$$

where, in this illustration, $t' = 5$. There is no ς_{i3} term ($i = 0, 1, 2$) because the misinformation model assumes that if examinees do not know, they cannot eliminate any of the distractors. More generally,

$$p_{11} = \varsigma_{t'-1,t-1} + \sum_{j=0}^{t-1} \varsigma_{t'-1,j}/(t - j)t' + \sum_{j=0}^{t-1} \varsigma_{0j}/(t - j)t' \qquad [39]$$

Also

$$p_{k1} = p_{11} - \varsigma_{42} \quad (k = 2, \ldots, t') \qquad [40]$$

$$p_{12} = \varsigma_{41}/2t' + \varsigma_{40} \qquad [41]$$

$$p_{1m} = \sum_{j=0}^{m} \varsigma_{4j}/(t - j)t' \quad (m = 0, \ldots, t-2) . \qquad [42]$$

The remaining p_{ij} values can be determined in a similar manner. For the two items being used here

$$p_{2m} = \sum_{j=0}^{m} \varsigma_{0j}(t - j)t' \quad (m = 2, \ldots, t) \qquad [43]$$

and $p_{3m} = p_{2m}$.

The expressions for p_{4m} and p_{5m} involve the proportion of examinees who have misinformation on the first item. The necessary equations can be derived as was illustrated above. This in turn yields an estimate of τ which can be used to estimate the bounds on p_k.

Testing Whether Items are Equivalent or Hierarchically Related

The model described in this paper might also be useful when empirically checking the assumptions of other latent structure models. For example, Macready and Dayton (1977) and Wilcox (1977) propose models where it is assumed that pairs of equivalent items are available. Two items are defined to be equivalent if examinees either know both or neither one. When equivalent items are available, the proportion of examinees who know both can be estimated (assuming local independence). Macready and Dayton checked their model with a chi-square goodness-of-fit test, but this requires at least three items that are equivalent to one another. (When there are only two items, there are no degrees of freedom left.)

For illustrative purposes, assume t=3, and consider equations 8-16. If two items are equivalent, then

$$\varsigma_{21} = \varsigma_{20} = \varsigma_{12} = \varsigma_{02} = 0 \qquad [57] \rightarrow 44$$

$$p_{12} = p_{21} = p_{22} \qquad [58] \rightarrow 45$$

$$p_{13} = p_{23} \qquad [59] \rightarrow 46$$

and

$$p_{31} = p_{23} \qquad [60] \rightarrow 47$$

For $N \leq 50$, an exact test of these last three equalities can be made using the critical

values in Katti (1973) and Smith et al. (1979) (Note that the conditional distribution of multinomial random variables is multinomial.) For larger N, the usual chi-square test can be used. From Smith et al. (1979), a slight adjustment to the usual chi-square test appears to be useful. Finally, if one of these items is assumed to be hierarchically related to the other, again certain equalities must hold among equations 8-16, and this can again be tested (cf. White and Clark, 1973; Dayton and Macready, 1976).

A Concluding Remark

It should be stressed that ρ_k is of interest <u>after</u> it has been decided which items are to be included on a test. ρ_k is not intended to measure validity -- it is designed to measure the overall effectiveness of the dis-tractors that are being used. Put another way, ρ_k is not meant to be the one and only index for characterizing a test -- it is intended to be one of several indices that might be used. The reason for raising this issue is that a test constructor can ensure that ρ_k is large by using easy items. This is an improper procedure that misses the point of how ρ_k is to be used.

Table 1

Number of Examinees Requiring i Attempts on Item 1

and j Attempts on Item 2

		Number of Attempts on Item 2			
		1	2	3	Total
Number of Attempts on Item 1	1	179	26	14	219
	2	76	8	4	88
	3	53	13	4	70
	Total	308	47	22	377

References

Alam, K., and Mitra, A. Polarization test for the multinomial distribution. Journal of the American Statistical Association, 1981, 76, 107-109.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. Statistical inference under order restrictions. New York: Wiley, 1972.

Bergan, J.R., Cancelli, A.A., and Luiten, J.W. Mastery assessment with latent class and quasi-independence models representing homogeneous item domains. Journal of Educational Statistics, 1980, 5, 65-81.

Bliss, L.B. A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 1980, 17, 147-153.

Brownless, V.T., and Keats, J.A. A retest method of studying partial knowledge and other factors influencing item response. Psychometrika, 1958, 23, 67-73.

Cross, L.H., and Frary, R.B. An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

Dayton, C.M., and Macready, G.B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Dayton, C.M., and Macready, G.B. A scaling model with response errors and intrinsically unscalable respondents. Psychometrika, 1980, 45, 343-356.

Katti, S. K. Exact distributuon for the chi-square test in the one way table. Communications in Statistics, 1973, 2, 435-447.

Knapp, T.R. The reliability of a dichotomous test-item: A 'correlationless' approach. Journal of Educational Measurement, 1977, 14, 237-252.

Macready, G.B., and Dayton, C.M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Macready, G.B., and Dayton, C.M. The nature and use of state mastery models. Applied Psychological Measurement, 1980, 4, 493-516.

Marks, E., and Noll, G.A. Procedures and criteria for evaluating reading and listening comprehension tests. Educational and Psychological Measurement, 1967, 27, 335-348.

Marshall, A., and Olkin, I. Inequalities: Theory of majorization and its applications. New York: Academic Press, 1979.

Molenaar, I. On Wilcox's latent structure model for guessing. British Journal of Mathematical and Statistical Psychology, 1981, 34, 224-228.

Pearson, K. Tables of the incomplete beta function. Cambridge: University Press, 1968.

Robertson, T. Testing for and against an order restriction on multinomial parameters. Journal of the American Statistical Association, 1978, 73, 197-202.

Sathe, Y.S. Pradhan, M., and Shah, S.P. Inequalities for the probability of the occurrence of at least m out of n events. Journal of Applied Probability, 1980, 17, 1127-1132.

Smith, P. J., Rae, D. S., Manderscheid, R. W., & Silbergeld, S. Exact and approximate distributions of the chi-square statistic for equiprobability. Communications in Statistics - Simulation and Computation, 1979, B8, 131-149.

Weitzman, R.A. Ideal multiple-choice items. Journal of the American Statistical Association, 1970, 65, 71-89.

White, R.T., & Clark, R.M. A test of inclusion which allows for errors of measurement. Psychometrika, 1973, 38, 77-86.

Wilcox, R.R. New methods for studying stability. In C.W. Harris, A. Pearlman, and R. Wilcox, Achievement test items: methods of study. CSE Monograph No. 6, Los Angeles: Center for the Study of Evaluation, University of California, 1977. (a)

Wilcox, R.R. New methods for studying equivalence. In C.W. Harris, A. Pearlman, and R. Wilcox, Achievement test items: Methods of study. CSE Monograph No. 6; Los Angeles: Center for the Study of Evaluation, University of California, 1977. (b)

Wilcox, R.R. Solving measurement problems with an answer-until-correct scoring procedure. Applied Psychological Measurement, 1981, 5, 399-414. (a)

Wilcox, R.R. Recent advances in measuring achievement: A response to Molenaar. British Journal of Mathematical and Statistical Psychology, 1981, 34, 229-237. (b)

Wilcox, R.R. Some empirical and theoretical results on an answer-until-correct scoring procedure. British Journal of Mathematical and Statistical Psychology, 1982, in press. (a)

Wilcox, R.R. Some new results on an answer-until-correct scoring procedure. Journal of Educational Measurement, 1982, 19, 67-74 (b)

Wilcox, R.R. Using results on k out of n system reliability to study and characterize tests. Educational and Psychological Measurement, in press. (b)

Wilcox, R.R. Determining the length of multiple-choice criterion-referenced tests when an answer-until-correct scoring procedure is used. Educational and Psychological Measurement, in press. (c)

AN APPROXIMATION OF THE K OUT N RELIABILITY OF A TEST, AND A SCORING PROCEDURE FOR DETERMINING WHICH ITEMS AN EXAMINEE KNOWS

Rand R. Wilcox
Department of Psychology
University of Southern California
and
Center for the Study of Evaluation
University of California, Los Angeles

ABSTRACT

Consider any scoring procedure for determining whether an examinee knows the answer to a test item. Let $x_i=1$ if a correct decision is made about whether the examinee knows the ith item; otherwise $x_i=0$. The k out of n reliability of a test is $\rho_k=Pr(\Sigma\ x_i \geq k)$. That is, ρ_k is the probability of making at least k correct decisions for a typical (randomly sampled) examinee. This paper proposes an approximation of ρ_k that can be estimated with an answer-until-correct test. The paper also suggests a scoring procedure that might be used when ρ_k is judged to be too small under a conventional scoring rule where it is decided an examinee knows if and only if the correct response is given.

Consider an n-item multiple-choice test, and suppose that every examinee can be described as either knowing or not knowing the correct response. In some situations, particularly with respect to some instructional program, the goal of a test might be to determine how many of the n items an examinee actually knows; in terms of diagnosis, it may even be desirable to determine which specific items an examinee knows or does not know. Under a conventional scoring procedure, about the only scoring rule available is one where it is decided that an examinee knows if and only if a correct response is given. Obviously guessing will affect the accuracy of this rule. If it is assumed that examinees who know will always give the correct response, and if most examinees really do know the correct response, then of course guessing has little impact on the accuracy of the test or the effectiveness of the distractors in terms of the typical examinee. However, if ς is the proportion of examinees who know the answer to an item, then as ς decreases, the importance of having effective distractors increases in order to avoid incorrect decisions about whether an examinee knows.

Guessing can seriously affect various other measurement problems as well (e.g., Weitzman, 1970; van den Brink and Koele, 1980; Wilcox, 1980, 1982c; Ashler, 1979). For example, when estimating the biserial correlation coefficient, guessing can substantially affect the results (Ashler, 1979). Ashler gives a method of correcting the estimate for the effects of guessing, but it requires a procedure for determining which items an examinee really knows. The conventional rule is to decide an examinee knows if and only if the correct response is given, but this can be unsatisfactory.

Suppose, for example, ζ=.5, and the probability of a correct response, given that the examinee does not know, is 1/3. Then 1/6 of the examinees would be misclassified. The extreme case is where none of the examinees know, in which case 1/4 would be incorrectly judged as knowing the correct response.

As another example, suppose an investigator wants to determine whether the proportion of examinees who know an item is relatively large. In order to ensure a reasonably high probability of a correct decision about this proportion, it follows from Wilcox (1980) that it might be necessary to sample ten, perhaps even forty times as many examinees as would be required if guessing did not exist.

For a specific examinee taking a test, let x_i=1 if a correct decision is made about whether the answer to the ith item is known; otherwise x_i=0. For an examinee randomly sampled from the population of potential examinees, let

$$\rho_k = Pr(\sum x_i \geq k).$$

This is just the probability of making at least k correct decisions among the n items for a randomly sampled examinee; ρ_k is called the k out of n reliability of a test.

Suppose every item has t alternatives. One approach to designing a reasonably accurate test is to assume random guessing, and then choose t so that ρ_k is reasonably close to one. If x_i is independent of x_j for all $i \neq j$, then ρ_k is easily calculated on a computer. Unfortunately, there are three serious problems with this approach. First, there is considerable empirical evidence that guessing is seldom at random (Coombs et al., 1956; Bliss, 1980; Cross & Frary, 1977; Wilcox, 1982a,

1982b). Second, even if guessing is at random, some situations will require more alternatives than is practical in order for ρ_k to be close to one (Wilcox, 1982c). Finally, there is no particular reason for assuming x_i independent of x_j, $i \neq j$, or to believe that such an assumption will give a good approximation of ρ_k. If $\text{cov}(x_i, x_j) \neq 0$, bounds on ρ_k are available (Wilcox, 1982c, in press a), but point estimates do not exist.

One goal in this paper is to suggest an approximation of ρ_k that can be estimated with an answer-until-correct test. Another and perhaps more important goal is to describe a scoring procedure that might be used when the estimate of ρ_k is judged to be too small under a conventional scoring rule. The new rule is based on a recently proposed latent structure model for test items. Included are some results on how to test whether this model is consistent with observed test scores.

2. An Approximation of ρ_k

Let $y=(y_1,\ldots,y_n)$ be any vector of length n where $y_i=0$ or 1, and let $f(y)$ be the probability density function of y. Bahadur (1961) shows that $f(y)$ can be written as

$$f(y) = f_1(y)h(y)$$

where

$$f_1(y)= \prod_{i=1}^{n} \alpha_i^{y_i} (1- \alpha_i)^{1-y_i}$$

$$\alpha_i=Pr(y_i=1)$$

$$h(y)=1 + \sum_{i<j} r_{ij} z_i z_j + \sum_{i<j<m} r_{ijm} z_i z_j z_m + \cdots + r_{12\ldots n} z_1 \cdots z_n$$

$$z_i = (y_i - \alpha_i)/[\alpha_i(1-\alpha_i)]^{\frac{1}{2}}$$

$$r_{ij} = E(z_i z_j)$$

$$r_{ijm} = E(z_i z_j z_m)$$

$$r_{12\ldots n} = E(z_1 z_2 \ldots z_n)$$

An mth order Bahadur approximation of f is one where the first m summations are used in the expression for h. Several authors have used a second order approximation when investigating problems in discrete discriminate analysis (e.g., Dillon & Goldstein, 1978; Gilbert, 1968; Moore, 1973). In this case f(y) is approximated with

$$g(y) = f_1(y)\left[1 + \sum_{i<j} r_{ij} z_i z_j\right] \qquad (2.1)$$

Other approximations have been proposed, but as will become evident, (2.1) is particularly convenient for the situation at hand.

Occasionally (2.1) will not be a probability function. In particular, it may be that $g(y) < 0$ for some vectors y. In this paper, whenever this occurred, g(y) was assumed to be zero, but the g(y) values were not re-scaled so that they sum to one.

Bahadur (1961) discusses how to assess the goodness of fit of the approximation. Here, however, interest is in approximating ρ_k. Note that for a random vector y, ρ_k can be written as

$$\sum_{y:S \geq k} f(y) \qquad (2.2)$$

where $S = \sum y_i$ and the summation in (2.2) is over all vectors y such that $S \geq k$.

Of course, when approximating ρ_k, f(y) would be replaced by f(x) where the vector x indicates which items a correct decision is made about whether an examinee knows. To gain some insight into how well g(y) approximates ρ_k, assuming α_i and r_{ij} are known, we set n=5, k=4 and randomly chose values for the 2^5=32 probability cells. Next, ρ_k was evaluated with (2.2), and then it was approximated with $\tilde{\rho}_k$ where $\tilde{\rho}_k$ is given by (2.2) with f(y) replaced by g(y). This process was repeated 100 times yielding a wide range of values for ρ_k. The values for ρ_k and $\tilde{\rho}_k$ were rounded to the second decimal place after which it was found that 85% of the time, $|\rho_k-\tilde{\rho}_k|\leq .02$. For 5% of the approximations it was found that $|\rho_k-\tilde{\rho}_k|\geq .05$. For $|\rho_k-\tilde{\rho}_k|\leq .05$ it was also found that $\tilde{\rho}_k<\rho_k$. The poorest approximation was for a probability function where ρ_k =.365 and $\tilde{\rho}_k$=.232. Although hardly conclusive, these results suggest that $\tilde{\rho}_k$ is generally useful when approximating ρ_k, at least when n is small. For n large the test can be broken into subtests containing five items or less, and Bonferroni's inequality (e.g., Tong, 1980) can be applied. For example, suppose n=10. If for the first five items $\tilde{\rho}_4$=.95, and for the remaining five items $\tilde{\rho}_4$=.98, then for the entire test it is estimated that

$$\rho_8\geq 1-(1-.98)-(1-.95)=.93. \qquad (2.3)$$

Estimating $\tilde{\rho}_k$

There remains the problem of estimating $\tilde{\rho}_k$. What is needed is an estimate of the parameter r_{ij} in the expression for g(y). An estimate is available using a slight extension of the model in Wilcox (in press.a) which can be briefly summarized as follows. Assume that examinees take the test according to an answer-until-correct scoring procedure. That is,

they choose a response, and if it is wrong they choose another. This process continues until the correct response is selected. Administering such tests is easily accomplished with especially designed answer sheets that are available commercially.

Consider a specific item and let P_i be the probability that a randomly selected examinee gets the item correct on the ith attempt, $i=1,\ldots,t$ where t is the number of alternatives. Let ς_i be the proportion of examinees who can eliminate i distractors $(i=0,\ldots,t-1)$. It is assumed that for examinees who do not know, there is at least one effective distractor in which case ς_{t-1} is the proportion of examinees who know. It is also assumed that once examinees eliminate as many distractors as they can, they guess at random from among those alternatives that remain. It follows that

$$P_i = \sum_{j=0}^{t-i} \varsigma_j/(t-j) \qquad (i=1,\ldots,t) \tag{2.4}$$

and the model implies that

$$P_1 \geq P_2 \geq \ldots \geq P_t \tag{2.5}$$

which can be tested (Robertson, 1978). For empirical results in support of this model, see Wilcox (1982a, 1982b, in press b). In the few instances where (2.5) seems to be unreasonable, a misinformation model appears to explain the observed test scores. When (2.5) is assumed, the pool-adjacent violators algorithm (Barlow et al., 1972) yields a maximum likelihood estimate of the P_i's. These estimates in turn yield an estimate of the ς_i's.

For any pair of items, let P_{ij} be the probability of a correct on the ith attempt of the first and the jth attempt of the second, respectively.

and let ς_{ij} be the probability that a randomly chosen examinee can eliminate i distractors from the first, and j distractors from the second. Then $\varsigma_{t-1,\,t-1}$ is the proportion of examinees who know both. It is assumed that an examinee's guessing rate is independent over the items not known, and so

$$P_{km} = \sum_{i=0}^{t-k} \sum_{j=0}^{t-m} \varsigma_{ij} \,/\, [(t-i)(t-j)] \qquad (2.6)$$

If the second item has t' alternatives, $t \neq t'$, simply replace t with t' in the second summation. Testing certain implications of (2.6) is discussed below.

For the ith item on the test, let $\tau_i = E(x_i)$ be the probability of a correct decision about whether the examinee knows when a conventional scoring procedure is used. Thus, τ_i plays the role of α_i when approximating ρ_k. For an answer-until-correct test, a conventional rule means to decide an examinee knows if and only if the correct response is given on the first attempt. In this case (Wilcox, 1982a)

$$\tau_i = \varsigma_{t-1}{}^{+1} - p_1$$

$$= 1 - p_2$$

Thus, if for the ith item, c_j of N examinees get the correct response on the jth attempt under an answer-until-correct scoring procedure, then

$$\hat{\tau}_i = 1 - c_2/N$$

is an estimate of τ_i. If the c_j's are inconsistent with (2.5), apply

the pool-adjacent-violators algorithm (Barlow et al., 1972, pp. 13-16), as was previously mentioned.

In a similar manner, let $\tau_{ij} = Pr(x_i=1, x_j=1)$, i.e., τ_{ij} is the probability of a correct decision for both items i and j. For the conventional decision rule under an answer-until-correct model, it can be seen that

$$\tau_{ij} = \sum_{k=1}^{t} \sum_{m=1}^{t} q_{km}$$

where

$$q_{11} = \varsigma_{t-1,\, t-1}$$

$$q_{i1} = \sum_{k=0}^{t-i} \varsigma_{k,t-1}/(t-k) \qquad (i=2,\ldots,t)$$

$$q_{1j} = \sum_{k=0}^{t-j} \varsigma_{t-1,k}/(t-k) \qquad (j=2,\ldots,t)$$

$$q_{ij} = P_{ij} \qquad (i>1 \text{ and } j>1)$$

(Wilcox, in press a)

Thus, r_{ij}, z_i and z_j in equation (2.1) are easily determined. In particular,

$$r_{ij} = \frac{\tau_{ij} - \tau_i \tau_j}{[\tau_i \tau_j (1-\tau_i)(1-\tau_j)]^{\frac{1}{2}}}$$

where τ_i plays the role of α_i in the definition of z_i. But as noted in Wilcox (in press), the ς_{ij}'s in equation (2.6) are easily estimated, and these estimates yield an estimate of τ_{ij} which in turn gives an estimate of r_{ij}. Hence, $\tilde{\rho}_k$ can be estimated with equation (2.1) which gives an approximation of ρ_k.

Testing Certain Implications of the Model

For any pair of items, equation (2.6) implies that

$$P_{11} \geq P_{12} \geq \cdots \geq P_{1t'} \geq P_{2t'} \geq \cdots \geq P_{tt'} \qquad (2.7a)$$

$$P_{11} \geq P_{21} \geq \cdots \geq P_{t1} \geq P_{t2} \geq \cdots \geq P_{tt'} \qquad (2.7b)$$

$$P_{i1} \geq P_{i2} \geq \cdots \geq P_{it'} \qquad (i=2,\ldots,t-1) \qquad (2.7c)$$

and

$$P_{1j} \geq P_{2j} \geq \cdots \geq P_{tj} \qquad (j=2,\ldots,t'-1) \qquad (2.7d)$$

where as before, t and t' are the number of alternatives for the first and second items, respectively. A few other inequalities are implied if the ζ_{ij}'s are assumed to be probabilities, but these have not been derived.

Experience with real data suggests that when observed scores are consistent with (2.5), the inequalities in (2.7) will also hold. If some of the observed proportions are inconsistent with (2.7), maximum likelihood estimates can be obtained when the model is assumed to be true by applying the minimax order algorithm in Barlow et al. (1972).

Robertson (1978) includes some asymptotic results on testing (2.7). At the moment, however, his proposed procedure can not be applied because certain constants (the $P_q(\ell,k)$'s in Robertson's notation) are not available. An alternative approach is to perform a separate test of the inequalities in (2.7d), one corresponding to every j, j=2,..., t'-1, then

perform a test of (2.7c), one for every i=2,...,t-1, then test (2.7b) and finally (2.7a). The total number of tests is m=t+t-2. If the critical value for every test is set at α/m, then from the Bonferroni inequality (e.g., Tong, 1980), the probability of a Type I error among the m tests is at most α.

Consider, for example, the inequalities in (2.7d) for j=2. That is, the goal is to test

$$H_0: \; p_{12} \geq p_{22} \geq p_{32} \geq \cdots \geq p_{t2} \tag{2.8}$$

Let λ be the likelihood ratio for testing (2.8) where the alternative hypothesis is no restriction on the proportions. From Robertson (1978, Theorem 2), the asymptotic null distribution of $T=-2 \ln \lambda$ is

$$\Pr(T>T_0) = \sum_{\ell=1}^{k-1} P(\ell,k) \Pr(\chi^2_{k-\ell} \geq T_0) \tag{2.9}$$

where $P(\ell,k)$ is the probability that the maximum likelihood estimate of $p_{12},\ldots,p_{t2}$ subject to (2.8) will have ℓ distinct values among the k parameters being estimated, and $\chi^2_{k-\ell}$ is a chi-square random variable with $k-\ell$ degrees of freedom. For (2.8), k=t. (As previously mentioned, the pool-adjacent-violators algorithm yields maximum likelihood estimates when (2.8) is assumed.) The constants $P(\ell,k)$ can be read from Table A.5 in Barlow et al. (1972).

Thus, in order for the m tests to have a critical level of at most α, choose T_0 so that (2.9) equals α/m, and reject H_0 if $T>T_0$. This process is repeated for the other inequalities to be tested, but note that k (the number of parameters being tested) will have a different value for (2.7a) and (2.7b).

To facilitate this procedure, critical values are reported in Table 1 for t=2(1)5, α=.1, .05, .01; and some appropriately chosen values for m. (Additional values for m were not used because for $t \leq 5$, these are the only values of m that will occur.)

As an illustration, suppose $t=t'=3$. Then there are m=4 sets of inequalities to be tested. If α=.05, then from (2.7a) there are k=5 parameters, and so T_0=10.81. For (2.7b) again k=5 and T_0=10.81. For (2.7c) there is only one set of inequalities which corresponds to i=2, t=k=3, and T_0=7.24. The same is true for (2.7d).

3. A Scoring Procedure for Tests

Consider a specific item on an n-item test. In contrast to most of the existing scoring procedures, the goal here is to minimize the expected number of examinees for whom an incorrect decision is made about whether they know the answer to the item. It is interesting to note that when items are scored right/wrong, this criterion can rule out the conventional rule where it is decided an examinee knows if and only if the correct response is given. The extreme case is where $\zeta_{t-1}=0$, i.e., none of the examinees know, in which case the optimal rule is to decide that an examinee does not know regardless of the response given. If β=Pr (correct | examinee does not know), it can be seen that if an item is scored right/wrong, and if $\beta > \zeta_{t-1}/(1-\zeta_{t-1})$ the optimal rule is to always decide that examinees do not know. If $\beta < \zeta_{t-1}/(1-\zeta_{t-1})$, use the conventional rule. From Copas (1974), this approach (in terms of parameters) is admissible.

These parameters can be estimated which yields an estimate of the optimal decision rule (e.g., Macready and Dayton, 1977). The goal here is to derive a decision rule based on an answer-until-correct scoring procedure. The advantage of this new approach is that it is not necessary to assume all n items are equivalent as was done in Macready and Dayton. (Two items are said to be equivalent if every examinee knows both or neither one.) The results in Macready and Dayton (1977) could be extended to the case of hierarchically related items by applying results in Dayton and Macready (1976), but here the goal is to derive a rule where no particular relationship is assumed among the items. However, the situation considered by Macready and Dayton (1977) has the advantage of allowing Pr (incorrect response | examinee knows) >0, while here this probability is assumed to be zero.

Consider the ith item on a test taken by a specific examinee, and let w_i=1 if it is decided the examinee knows; otherwise w_i=0. Consider the jth item on the test i≠j for the purpose of assisting in the decision about whether w_i should be 1 or 0. (The optimal choice for the second item will become evident.) It is assumed that items are administered according to an answer-until-correct scoring procedure. For a specific examinee, let v_i be the number of attempts needed to choose the correct response to the ith item. The decision rule to be considered is

$$w_i(v_j) = \begin{cases} 1, \text{ if } v_i < v_{0i} \text{ and } v_j \leq v_{0j} \\ 0, \text{ otherwise} \end{cases} \qquad (3.1)$$

where v_{0i} (=1 or 2) and v_{0j} ($1 \leq v_{0j} \leq t'$) are constants to be determined. Note that when $v_{0i}=2$ and $v_{0j}=1$, the rule is similar to the one in Macready and Dayton (1977). Also note that $v_{0j}=t'$ corresponds to the conventional decision rule where the information about the jth item plays no role in determining whether the examinee knows the ith. It is evident, therefore, that in terms of parameters, (3.1) always improves upon the conventional approach. The improvement actually achieved will of course vary. If ζ_{t-1} is close to one for every item, ρ_k will also be close to one under a conventional scoring rule, in which case there is little motivation for using (3.1). However, when ρ_k is unacceptably small, (3.1) can increase ρ_k by a substantial amount.

One problem is choosing the constants v_{0i} and v_{0j}. A solution is as follows. For a randomly sampled examinee responding to the ith and jth items, let P_{kml} be the probability of choosing the correct response on the kth attempt of the ith item, the mth attempt of the jth item, and making a correct decision under the rule (3.1). The probability of a correct decision for a randomly sampled examinee is

$$p_c = \sum_{k=1}^{t} \sum_{m=1}^{t'} p_{kml}$$

which is a function of v_{0i} and v_{0j}. Thus, the obvious choice for v_{0i} and v_{0j} is the one that maximizes p_c.

Let

$$q_{1j} = \sum_{k=0}^{t'-j} \zeta_{t-1,k}/(t'-k) \qquad (j=1,\ldots,t')$$

and

$$Q = \sum_{i=2}^{t} \sum_{j=1}^{t} P_{ij}$$

For $v_{0i}=2$ and any v_{0j}

$$P_c = Q + \sum_{k=1}^{v_{0j}} q_{1k} + \sum_{k=v_{0j}+1}^{t'} (p_{1k}-q_{1k}) \qquad (3.3)$$

When $v_{0j}=t'$, the second sum in (3.3) is taken to be zero. As for $v_{0i}=1$,

$$p_c = Q + \sum_{k=1}^{t'} (p_{1k}-q_{1k}). \qquad (3.4)$$

Thus, to determine the optimal choice for v_{0i} and v_{0j} in (3.1), simply evaluate p_c for every possible choice of v_{0i} and v_{0j}, and then set v_{0i} and v_{0j} equal to the values that maximize p_c. Of course, when making a decision about the ith item, this process can be repeated over the n-1 other items on the test. The item that maximizes p_c is the one that should be used when determining whether an examinee knows the ith item.

An Illustration

As a simple illustration, the optimal rule is estimated for two items used in Wilcox (1982a). The observed frequencies are shown in Table 2. Note that the observed frequencies already satisfy (2.7a)-(2.7d). For the first item the estimate of τ, the probability of correctly determining whether a randomly sampled examinee knows, is $\hat{\tau}_1=(236-71)/236=.699$. For the second items it is $\hat{\tau}_2=.78$.

Suppose the second item is used to help determine whether an examinee knows the first. Let $v_{01}=2$ and $v_{02}=1$. Thus, a correct response must be

given on the first attempt of both items in order to decide that an examinee knows. Note that

$$Q=1-\sum_{j=1}^{4} p_{1j},$$

and so Q is estimated to be .513.

The easiest way to estimate the ζ_{ij}'s is to start with $p_{tt'}$. From (2.6) with $t=t'=4$,

$$p_{44}=\zeta_{44}/16$$

and so from Table 2, $\hat{\zeta}_{44}=0$.

Next consider

$$p_{43}=\zeta_{01}/12+\zeta_{44}/16.$$

and so $\hat{\zeta}_{01}=.048$. Eventually this process yields estimates of all the ζ_{ij}'s which in turn yields estimates of q_{1j}, $j=1,\ldots,4$. The estimates turn out to be $\hat{q}_{11}=.109$, $\hat{q}_{12}=.012$, $\hat{q}_{13}=.012$, and $\hat{q}_{14}=0$. Thus, the estimate of P_c is $.513+.109+(.089-.012)+(.042-.012)+(.013-0.0)=.742$.

If instead $v_{01}=1$, so that it is always decided an examinee does not know, regardless of the observed response, P_c is just one minus the proportion of examinees who know. From (2.4), $\zeta_{t-1}=p_1-p_2$, so the estimate of P_c is $1-.187=.813$. This is a substantial increase in accuracy over the conventional rule.

Determining ρ_k Under the New Scoring Procedure

It is evident that the scoring procedure represented by (3.1) improves upon the conventional scoring procedure, but when v_{0j} in (3.1) is less than t', the method already described for determining ρ_k will in general be inadequate. The reason is that to determine ρ_k, τ_{ij}

(the joint probability of making a correct decision about the ith and jth item) must be known. (See Section 2.) But when $v_{0j}<t$, τ_{ij} may depend on two other items, say items k and m. That is, information on the kth item and mth item will be used to determine whether the examinee knows the ith and jth items respectively. Hence, (2.6) is no longer adequate for determining ρ_k.

One solution might be to extend (2.6) to include four items. In theory the parameters could be estimated under the resulting inequalities by applying the minimax order algorithm. However, writing an appropriate computer program that is valid for $t\leq 5$ will be a relatively involved task.

Another and perhaps more practical approach might be to restrict the decision rule so that if the response to the jth item is used in the decision about whether an examinee knows the ith item, then the response to the ith will be used in deciding about the jth. An advantage of this approach is that it simplifies the process of choosing a decision rule by reducing the number of pairs of items that are considered. A second advantage is that an approximation of ρ_k can be made using the results in section 2. A disadvantage is that by restricting the class of decision rules, the potential increase in ρ_k (over what it is under a conventional scoring rule) is reduced. Perhaps this is not a serious problem; at the moment it is impossible to say.

An approach to choosing a scoring rule might be as follows: First estimate ρ_k under conventional scoring rule. If it is judged to be too small, choose a decision rule from among the rules described in the preceding paragraph and then estimate ρ_k in the manner indicated below. If ρ_k is still too small, choose a decision rule from among the broader class

of rules described in the preceding subsection. In this case, however, an approximation of ρ_k is no longer available for the reasons just given.

Suppose that if the jth item is chosen to aid in the decision about the ith, then the ith item is used in the decision rule for the jth. What is needed in order to approximate ρ_k is an expression for the joint probability of making a correct decision for both items. Accordingly, consider any two items, and let $u_1(k,m)=1$ if it is decided that an examinee knows the first item if the correct response is given on the kth attempt of the first item, and the mth attempt of the second; otherwise $u_1(k,m)=0$. Similarly, $u_2(k,m)=1$ if it is decided that an examinee knows the second item if the correct response is given on the kth attempt of the first and the mth attempt of the second; otherwise $u_2(k,m)=0$. Let

$$s_{1i}(k,m)=\begin{cases}1, \text{ if } u_1(k,m)=1 \text{ and } i=t-1, \text{ or if} \\ \quad u_1(k,m)=0 \text{ and } i<t-1 \\ 0, \text{ otherwise}\end{cases}$$

and

$$s_{2j}(k,m)=\begin{cases}1, \text{ if } u_2(k,m)=1 \text{ and } j=t'-1, \text{ or if} \\ \quad u_2(k,m)=0 \text{ and } j<t'-1 \\ 0, \text{ otherwise.}\end{cases}$$

Recall that the probability of getting the correct response on the kth attempt of the first item and the mth attempt of the second is given by (2.6). From this expression it can be seen that the joint probability of k attempts on the first item, m attempts on the second, and a correct

decision on both items is

$$\gamma_{km} = \sum_{i=0}^{t-k} \sum_{j=0}^{t'-m} s_{1i}(k,m)s_{2j}(k,m)\zeta_{ij}/[(t-i)(t'-j)]$$

Thus, for a randomly sampled examinee, the joint probability of a correct decision for both items, say items i and j, is

$$\tau_{ij} = \sum_{k=1}^{t} \sum_{m=1}^{t'} \gamma_{km}$$

The joint probability of a correct decision about the first item, k attempts on the first and m attempts on the second is

$$\psi_{km} = \sum_{i=0}^{t} \sum_{j=0}^{t'} s_{1i}(k,m)\ \zeta_{ij}/[(t-i)(t'-j)]$$

The corresponding probability for the second item is

$$\eta_{km} = \sum_{i=0}^{t} \sum_{j=0}^{t'} s_{2j}(k,m)\zeta_{ij}/[(t-i)(t'-j)]$$

Thus, τ_i, the probability of a correct decision about the ith item on a test (using the jth item in (3.1)) for a randomly sampled examinee, is

$$\tau_i = \sum_{k=1}^{t} \sum_{m=1}^{t'} \psi_{km}$$

Similarly, for the second item, item j,

$$\tau_j = \sum_{k=1}^{t} \sum_{m=1}^{t'} \eta_{km}$$

Hence, ρ_k can be approximated as described in section 2.

Concluding Remarks

Virtually all of the results on the proposed scoring rule have been in terms of parameters. These parameters are not known, but they are easily estimated. The question arises as to the sampling effects on estimating the approximation of ρ_k, and on estimating the optimal decision rule for determining whether an examinee knows the correct response. In some instances, a large number of examinees will be available, and so very accurate estimates of the parameters can be obtained. This is the case for certain testing firms where literally thousands of examinees take the same test. When the number of examinees is small, however, sampling fluctuations need to be taken into account; this problem is currently being investigated.

Another important feature of the proposed scoring rule is that the decision about whether an examinee knows an item is a function of the responses given by the other examinees. If the goal is to minimize the number of examinees for whom an incorrect decision is made, there is no problem. However, in some instances, this feature might be objectionable. Suppose, for example, an examinee takes a test to determine whether a high school diploma will be received. It is possible for an examinee to fail because of how other examinees perform on the test even though the examinee in question deserves to pass. If this type of error is highly objectionable, perhaps the proposed scoring rule should be used only in diagnostic situations where the goal is to determine how many

items an examinee actually knows, or which specific items are not known.

A technical point that should be mentioned is that a few of the $\hat{\zeta}_{ij}$'s were slightly negative in which case $\hat{\zeta}_{ij}$ was set equal to zero. As a result, the $\hat{\zeta}_{ij}$'s sum to .997 rather than one as they should. The problem is that equations (2.7a)-(2.7d) are necessary but not sufficient conditions for the model to hold. For example, these inequalities do not guarantee that ζ_{12} will be positive.

Despite these difficulties, there will be situations where correcting for guessing can be important. Some examples were given at the beginning of the paper. Even if a conventional scoring procedure is to be used in operational versions of a test, it might be important to first estimate the effects of guessing using an answer-until-correct scoring procedure.

Many scoring rules have been proposed that are based on various criteria. If a particular criterion is deemed important, of course the corresponding scoring rule should be considered. The point is that most of these rules are not based on the goal of determining how many items an examinee knows, or which specific skills an examinee has failed to learn. Moreover, typical rules usually ignore guessing or assume guessing is at random. Thus, the results reported here might be useful in certain situations.

TABLE 1

Critical Values T_0 for the Bonferroni Test of Equations (2.7)

k	m	α: .1	.05	.01
3	4	5.90	7.24	10.38
3	5	6.33	7.67	10.81
3	6	6.68	8.03	11.17
4	3	7.03	8.49	11.86
4	4	7.64	9.10	12.46
4	5	8.11	9.56	12.92
4	6	8.49	9.95	13.30
5	4	9.25	10.81	14.36
5	5	9.75	11.31	14.85
5	6	10.16	11.71	15.25
5	7	10.51	12.05	15.58
5	8	10.81	12.35	15.87
6	5	11.32	12.96	16.67
7	6	13.29	15.00	18.85
8	7	15.18	16.95	20.93
9	8	17.02	18.84	22.94

Table 2

OBSERVED FREQUENCIES FOR TWO ITEMS ADMINISTERED UNDER AN ANSWER-UNTIL-CORRECT SCORING PROCEDURE

Number of Attempts for the First Item	Number of Attempts for the Second Item: 1	2	3	4	
1	81	21	10	3	115
2	44	18	6	3	71
3	20	7	5	1	33
4	10	6	1	0	17
	155	52	22	7	236

References

Ashler, D. Biserial estimators in the presence of guessing. Journal of Educational Statistics, 1979, 4, 325-355.

Bahadur, R. R. A representation of the joint distribution of responses to n dichotomous items. In H. Solomon (Ed.) Studies in Item Analysis and Prediction. Stanford: Stanford University Press.

Barlow, R., Bartholomew, D., Bremner, J., & Brunk, H. Statistical inference under order restrictions. New York: Wiley, 1972.

Bliss, L. B. A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 1980, 17, 147-153.

Coombs, C. H., Milholland, J. E., & Womer, F. B. The assessment of partial information. Educational and Pyschological Measurement, 1956, 16, 13-37.

Copas, J. B. On symmetric compound decision rules for dichotomies. Annals of Statistics, 1974, 2, 199-204.

Cross, L. H., & Frary, R. B. An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

Dayton, C. M., & Macready, G. B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Dillon, W. R., & Goldstein, M. On the performance of some multinomial classification rules. Journal of the American Statistical Association, 1978, 73, 305-313.

Gilbert, E. S. On discrimination using qualitative variables. Journal of the American Statistical Association, 1968, 63, 1399-1412.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Moore II, D. H. Evaluation of five discrimination procedures for binary variables. Journal of the American Statistical Association, 1973, 68, 399-404.

Robertson, T. Testing for and against an order restriction on multinomial parameters. Journal of the American Statistical Association, 1978, 73, 197-202.

Tong, Y. L. Probability inequalities in multivariate distributions.

van den Brink, W. P., & Koele, P. Item sampling, guessing and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1980, 33, 104-108.

Weitzman, R. A. Ideal multiple-choice items. Journal of the American Statistical Association, 1970, 65, 71-89.

Wilcox, R. R. Determining the length of a criterion-referenced test. Applied Psychological Measurement, 1980, 4, 425-446.

Wilcox, R. R. Some empirical and theoretical results on an answer-until-correct scoring procedure. British Journal of Mathematical and Statistical Psychology, 1982, 35, 57-70. (a)

Wilcox, R. R. Some new results on an answer-until-correct scoring procedure. Journal of Educational Measurement, 1982, 19, 67-74. (b)

Wilcox, R. R. Using results on k out of n system reliability to study and characterize tests. Educational and Psychological Measurement, 1982, 42, 153-165. (c)

Wilcox, R. R. Bounds on the k out of n reliability of a test, and an exact test for hierarchically related items. Applied Psychological Measurement, in press. (a)

Wilcox, R. R. How do examinees behave when taking multiple-choice tests? Applied Psychological Measurement, in press. (b)

HOW DO EXAMINEES BEHAVE WHEN TAKING MULTIPLE CHOICE TESTS?

Rand R. Wilcox

Center for the Study of Evaluation
University of California, Los Angeles

Horst (1933) assumed that when examinees respond to a multiple-choice test item, they eliminate as many distractors as possible, and guess at random from among those that remain. More recently, Wilcox (1981) proposed a latent structure model for achievement test items that was based on this assumption and which solves various measurement problems. (See also, Wilcox, 1982a, 1982b.)

Suppose an item is administered according to an answer-until-correct (AUC) scoring procedure. That is, examinees chose a response, and they are told whether it is correct. If incorrect they choose another response, and this process continues until the correct response is selected. Now consider two specific distractors. If Horst's assumption is true, then among the examinees choosing these two distractors, the order in which they are chosen should be at random. Of course for 3 distractors the same conclusion holds, only now there are 6 patterns of responses rather than 2. An empirical investigation of this implication is described below.

As was done on previous tests, the final examination for students enrolled in an introductory psychology course was administered according to an AUC scoring procedure. For 26 items, examinees were asked to record the order in which they chose their responses. Bonus points were given to those examinees complying with this request. There were 236 examinees who took the first 13 items, and 237 examinees took the remaining 13. All items had 4 alternatives.

For any two distractors, the null hypothesis of random order in responses can be tested with the usual sign test. Among the examinees

choosing all three distractors, the chi-square test given by

[Insert Equation 1 here]

was used where x_i is the number of examinees choosing the ith response pattern, and N is the number of examinees choosing all three distractors. Some exact critical values are given by Katti (1973), and Smith et al., (1979) and they were used whenever possible. For larger values of N, the adjusted chi-square test was used (Smith et al., 1979).

For each item, the responses to all pairs of distractors were tabulated. For N<5, no test was made because it is impossible to reject the null hypothesis at the .1 level. For the first test form, 29 tests were made and the hypothesis of random choices was rejected three times at the .1 level. For the second test form, 25 tests were performed, and again H_0 was rejected 3 times.

Next, an analysis was performed on those responses where all three distractors were chosen. Again no test was made for N<5. The largest value for N was 75.

At the .1 level, H_0 was rejected for 5 of the 12 items on the first test form, and for the second test form the rejection rate was 3 of 11.

The question remains as to the relative extent to which responses are not at random when H_0 is rejected. For the case where all three distractors were chosen, this quantity was measured with

[Insert Equation 2 here]

where χ^2_{max} and χ^2_{min} are the maximum and minimum possible values of χ^2. From Smith et al. (1979), $\chi^2_{max} = 5N$, and χ^2_{min} is given by Dahiya (1971). The quantity w has a value between 0 and 1 inclusive. The closer w is to one, the more unequal are the cell probabilities in a multinomial distribution.

Marshall and Olkin (1979) suggest that when measuring inequality, a certain class of functions (called Schur functions) should be used. Writing Equation 1 as a function of Σx_i^2 (Dahiya, 1971) and noting that Σx_i^2 is just Simpson's measure of diversity, it follows from results in Marshall and Olkin (1979) that w is a Schur function.

Note that using the w statistic is similar to using Hays' ω^2 (Hays, 1973). That is, rejecting the null hypothesis does not indicate the extent to which the cell probabilities are unequal. It may be, for example, that the cell probabilities are not equal, but that for practical purposes they are nearly the same in value.

For the first test form where H_0 was rejected, the w values were found to be .074, .183, .286, .167, and .137. For the second test form they were .098, .125 and .133. Thus, even when H_0 is rejected, Horst's assumption appears to be a tolerable approximation of reality in most cases. Of course there will probably be items where this assumption is grossly inadequate. In this case the measurement procedures proposed by Wilcox may be totally inappropriate.

References

Dahiya, R. D. On the Pearson chi-squared goodness-of-fit test statistic. Biometrika, 1971, 58, 685-686.

Hays, W. Statistics for the Social Sciences. New York: Holt, Rinehart and Winston, 1973.

Horst, P. The difficulty of a multiple-choice test item. Journal of Educational Psychology, 1933, 24, 229-232.

Katti, S. K. Exact distribution for the chi square test in the one way table. Communications in Statistics, 1973, 2, 435-447.

Marshall, A., & Olkin, I. Inequalities: Theory of majorization and its applications. New York: Academic Press, 1979.

Smith, P. J., Rae, D. S., Manderscheid, R. W., & Silbergeld, S. Exact and approximate distributions of the chi-square statistic for equiprobability. Communications in Statistics — Simulation and Computation, 1979, B8, 131-149.

Wilcox, R. R. Solving measurement problems with an answer-until-correct scoring procedure. Applied Psychological Measurement, 1981, 5, 399-414.

Wilcox, R. R. Some empirical and theoretical results on an answer-until-correct scoring procedure. Bristish Journal of Mathematical and Statistical Psychology, 1982a, 35, in press.

Wilcox, R. R. Some new results on an answer-until-correct scoring procedure. Journal of Educational Measurement, 1982b, in press.

$$X^2=\Sigma(x_i-N/6)^2/(N/6) \qquad [1]$$

$$w=(X^2-X^2_{min})/(X^2_{max} - X^2_{min}) \qquad [2]$$